Development and Validation of the Light and Spectroscopy Concept Inventory

by **Erin M. Bardar**
Boston University and TERC
Edward E. Prather
University of Arizona
Kenneth Brecher
Boston University
Timothy F. Slater
University of Arizona

Abstract

This article describes the development and validation of the Light and Spectroscopy Concept Inventory (LSCI), a 26-item diagnostic test designed (1) to measure students' conceptual understanding of topics related to light and spectroscopy, and (2) to evaluate the effectiveness of instructional interventions in promoting meaningful learning gains in an introductory college astronomy course. We also present the final field- tested version of the LSCI for general use by the astronomy education community.

1. INTRODUCTION

In our previous paper (Bardar (Weeks) et al. 2005), we established the motivation for developing a single standardized assessment instrument for the evaluation of introductory college astronomy teaching and learning, centered on concepts related to light and spectroscopy. In this article, we describe in detail the process by which this test, the Light and Spectroscopy Concept Inventory (LSCI), was developed. We also make the test available for general use within the astronomy education community.

The construction of the LSCI followed rigorous test development guidelines defined by classical test theory, whose principles were derived expressly for education and social science research. The list below is an adaptation of the steps of test construction as discussed by Crocker and Algina (1986).

Process of test construction:

1. Identify the primary purpose for which test scores will be used.
2. Define the concept domain to be addressed by the test.
3. Construct and review an initial pool of items.
4. Hold preliminary item tryouts and revise as necessary.
5. Field-test items with a large sample representative of the population for whom the test is intended.
6. Determine statistical properties of item scores and, when appropriate, eliminate items that do not meet pre-established criteria.
7. Conduct reliability and validity studies for the final form of the test.

2. PURPOSE AND SPECIFICATIONS

As discussed in our previous paper (Bardar (Weeks) et al. 2005), the purpose of developing a Light and Spectroscopy Concept Inventory for the introductory college astronomy survey course was twofold: (1) to produce a standardized assessment instrument for evaluating individual students' conceptual understanding of topics central to the course, and (2) to produce a tool for evaluating and comparing the relative effectiveness of various instructional methods in promoting conceptual change.

Through an extensive review of journal articles and surveys of faculty, syllabi, and textbooks, we came to the conclusion that the electromagnetic spectrum and the nature of light are widely acknowledged within the astronomy education community as the most taught and most important topics in introductory astronomy courses (Slater & Adams 2003; Slater et al. 2001; Zeilik & Morris-Dueer 2005) and as topics with which students struggle (Brecher 1991; Zeilik, Schau, & Mattern 1998). These topics therefore present themselves as a natural choice for a central theme around which to develop a concept inventory that would allow astronomy instructors to compare the success of the teaching and learning that takes place in the widest possible range of introductory courses.

The concept domain of the LSCI, shown below, was chosen to reflect the most commonly taught concepts addressed by the majority of courses within the astronomical community.

Concepts addressed by the LSCI:

- The nature of the electromagnetic spectrum, including the interrelationships of wavelength, frequency, energy, and speed
- Interpretation of Doppler shift as an indication of motion rather than color of an object
- The correlation between peak wavelength and temperature of a blackbody radiator
- Relationships between luminosity, temperature, and surface area of a blackbody radiator
- The connection between spectral features and underlying physical processes

3. ITEM DEVELOPMENT

To minimize the time and effort required to administer and score the test, a multiple-choice format was chosen for the LSCI. Items were generated based on prior research into students' understanding of light and quantum phenomena (Ambrose, Heron, et al. 1999; Ambrose, Shaffer, et al. 1999; Comins 2001; Langley, Ronen, & Eylon 1997; Lee 2002; Wosilait et al. 1999); the combined teaching experience of the

authors; and the results of in-depth student interviews. These multiple-choice items were configured to consist of a question stem followed by three to five answer choices: the scientifically correct answer and known alternate conceptions, or "distracters." Although most standard scannable forms have spaces for five response choices, there is no established rule of test construction theory that requires five choices to be supplied. Good distracters should appear incorrect to someone who fully understands the concept addressed by the item but should also appear reasonable to someone who does not understand the concept, therefore making them attractive response options. Nothing is gained through the inclusion of additional nonsensical distracters (Schwab 1963; Slater & Adams 2003). Distracters were written in the students' natural language to reflect commonly held misconceptions as revealed by previous research studies and responses to interview questions.

Standardized multiple-choice tests have often been criticized for testing only trivial, low-level thinking (Haladyna 1994). However, by consulting results from qualitative research studies, test developers are able to include attractive distracters written in the natural language of students, thus ensuring that the test elicits known misconceptions and reasoning difficulties rather than just factual recall (Tamir 1971). Answer choices that force students to confront known difficulties also allow instructors to more accurately determine deficiencies in understanding and decipher whether students choose a correct answer because they truly understand the concept, or whether they choose a wrong answer based on those known difficulties (Hufnagel 2002). Such distracter-driven multiple-choice instruments have been shown to be valuable tools for providing formative feedback to curriculum developers and for summative evaluation of experiments with new instructional materials (Sadler 1998).

For the LSCI, item stems were constructed to allow the correct answer to be determined even if the answer choices were not provided. Every effort was made to ensure that each item was as free of scientific jargon as possible. Questions were also constructed such that each item addressed a single concept and had answer choices of similar length and complexity.

Once an initial pool of questions was assembled, those items were distributed to other astronomy educators and researchers for review. These colleagues were asked specifically to examine and critique questions for content accuracy, wording and grammar, level of difficulty, and other technical flaws. Where there was disagreement or concern, suggested changes were discussed, and questions were revised accordingly.

4. PRELIMINARY ITEM TRYOUTS

Preliminary LSCI item tryouts were conducted with 50 non–science major undergraduate students taking AS102, "The Astronomical Universe," at Boston University. The LSCI was administered as a pretest during the first week of class and as an end-of-semester posttest. The inventory was also used as the protocol for one-on-one postinstruction interviews. During the semester, students attended two 90-minute lectures and a 90-minute laboratory section per week, and two one-hour nighttime observing sessions. Quantitative data revealed several problematic items. Over 75% of students were able to correctly answer 4 of 26 items prior to instruction, indicating that the concepts addressed by those items were inappropriate for an instrument designed to evaluate the relative effectiveness of instructional practices. Poor postinstruction performance (negative gain score), combined with feedback provided in interviews, revealed that students found one additional item to be ambiguously worded. These problematic items were thoroughly revised or rewritten. The updated instrument was redistributed to colleagues for a second round of review. The 26 questions remaining after this review process became the version of the LSCI

administered in the nationwide field test described in the following section.

5. FIELD TEST

In the fall of 2005, a multi-institution field test was conducted with student examinees from 14 course sections at 11 colleges and universities, including Boston University. All participants were enrolled in a 100-level introductory astronomy course for non–science majors. Of all participants, nearly 86% were taking an astronomy course for the first time. Volunteer instructors for this field test were recruited at the January 2005 national meeting of the American Astronomical Society and through e-mail solicitation. Prospective participants were presented with an overview of the research goals of the project and the logistical procedures associated with the administration of the inventory. Instructors choosing to participate were provided with a master copy of the LSCI (for photocopying and distributing); Scantron answer sheets; detailed instructions for data collection and preservation of student anonymity; and a prepaid return envelope for completed answer sheets. Instructors were asked to collect LSCI test forms from their students to minimize the risk of posttest data contamination as a result of studying the test questions. Instructors were also asked to fill out and return a brief survey describing the nature of their course with respect to the inclusion of active engagement activities. Results of this survey indicated that 11 courses were "traditional," primarily lecture-based courses, incorporating minimal use of student-centered learning activities; the remaining three courses were labeled as "active engagement" courses for their significant use of peer instruction (Mazur 1997), Lecture Tutorials (Prather et al. 2004), or homelabs (Bardar 2006).

Boston University policy mandates that all research involving human subjects or data is subject to approval by the university's Institutional Review Board (IRB). Before commencing LSCI data collection, a detailed overview of the project and student involvement was submitted to the IRB. Included in this proposal were student consent forms for participation in interviews and written data collection, interview protocols, and confidentiality guidelines. Numerical codes were used in place of names on all written data, and interview audio recordings were conducted anonymously and kept in a locked office to be destroyed five years after the completion of the research project. The course instructor had no access to the data prior to the conclusion of the course. No data obtained through interviews or course examinations were used for this research project without express written consent from the student. Students at other colleges and universities supplied only written responses to the LSCI and did so voluntarily and anonymously. Additionally, no instructor outside Boston University acted as a primary researcher for the project. Therefore, IRB approval was only required through Boston University and not through each college and university involved in the study.

Class sizes ranged from 7 to 68 students, for a combined total number, N, of 548 students preinstruction and 368 students postinstruction. The decrease in sample size was attributed to standard dropout rates over the course of a semester, absences on the day of posttesting, and the withdrawal of one instructor (two course sections) midway through the study. The minimum acceptable sample size for a full-scale field test is typically 5–10 times as many subjects as test items (Nunnally 1967). Therefore, for the 26-item LSCI, the minimum valid sample population for field testing would be 130–260 student examinees. This figure is significantly smaller than the number of students who participated in the field test, indicating that the sample size was more than adequate for obtaining viable data for statistical analysis. The results from this nationwide field test of the LSCI are described in Bardar (2006, 2007).

6. ITEM ANALYSIS: DIFFICULTY AND DISCRIMINATION

Following preliminary item tryouts and a large-scale field test, statistical properties of item scores were examined to determine which items, if any, were not functioning as intended. Two item characteristics, item difficulty and item discrimination, were considered in deciding which items of the LSCI should be retained for future administration of the LSCI, and which items should be amended or removed altogether from the test.

6.1 Difficulty

Item difficulty, p, is defined as the proportion of students answering that item correctly (Crocker & Algina 1986). Because the LSCI is intended to function as a tool to gauge the relative effectiveness of instructional treatments, it was decided that a wide range of item difficulties ($0.2 < p < 0.8$) was desirable. Items with unsatisfactory p-values were flagged for revision or elimination pending further investigation of the item. A high p-value may indicate that the item was too easy for the test's intended population and therefore may not be appropriate for inclusion in an inventory designed to elicit misconceptions and reasoning difficulties and to gauge the relative effectiveness of instructional interventions. A high p-value may also be an indicator of a poorly written item in which the correct answer was somehow obvious to test takers based on factors unrelated to the conceptual content of the item (i.e., one answer choice was significantly longer or shorter than the others or written in a much more sophisticated or "scientific-sounding" style). Similarly, a low p-value does not necessarily indicate a malfunctioning item. A good item can sometimes be answered incorrectly by a majority of students because it addresses a particularly deep-rooted misconception or reasoning difficulty that is not easily reversed by instruction. This is a likely occurrence with a research-based concept inventory. Based on postinstruction analysis, two items were flagged for having p-values less than 0.20 (fewer than 20% of students answered the items correctly). A third item was also selected for further review, with a p-value of 0.209, on the cusp of the desired range of difficulties.

6.2 Discrimination

Item discrimination, which measures how well an item differentiates between examinees who score relatively high or low on the entire inventory, is generally a more insightful statistic than item difficulty. The point biserial correlation coefficient, rho(pbis) (Crocker & Algina 1986), represents the correlation between a dichotomous variable (correct = 1; incorrect = 0) and a continuous variable—in this case, the item score and the overall test score. With mu(plus) = the mean test score for students who answered the question correctly; mu(x) = the mean test score for the entire test population; sigma(x) = the standard deviation of the full set of test scores; p = item difficulty, and q = (1-p), the point biserial correlation coefficient is calculated by the formula

$$\rho_{pbis} = \frac{(\mu_+ - \mu_x)}{\sigma_x} \sqrt{p/q}$$

Values of the point biserial coefficient can range from -1.00 to +1.00. A positive value of the point biserial coefficient indicates a positive correlation between the item score and the overall test score—that is, students scoring higher on the exam were more likely to answer that particular item correctly and the item successfully discriminated between high- and low-scoring students. Likewise, a negative point biserial coefficient signifies that high scorers are answering the item incorrectly more frequently than low scorers and may signify a problem with the question (i.e., miskeyed answer sheet, misleading question, or that the content was not adequately covered or taught in the course).

Point biserials within the range 0.30 to 0.70 are considered desirable (Allen & Yen 1979). However, items with correlation indices lower than 0.30 may be considered for retention if the point biserial is significantly greater than 0.00 (Crocker & Algina 1986; Ebel 1965). For a sample of at least 50 examinees, the minimum critical value of the point biserial coefficient is set at two standard deviations above 0.00, with the standard deviation calculated by

$$\sigma_\rho = \frac{1}{\sqrt{N-1}}$$

where N is the sample size. For the postinstruction field test sample of 368 students, the standard deviation of the point biserial coefficient was 0.052. Therefore, items with point biserial coefficients greater than 0.104 may be retained if other predetermined item criteria are satisfactorily met. Although all items were found to have point biserial indices above 0.104, four items had discrimination values between 0.104 and 0.30. These items, along with the items previously flagged for unsatisfactorily low p-values, were reviewed individually to determine whether they should be kept, revised, or eliminated for future administrations of the LSCI. Details of this item review can be found in Bardar (2006).

7. ESTABLISHING RELIABILITY AND VALIDITY

To convince potential users that the LSCI would have the ability to measure their students' conceptual understanding of light and spectroscopy, and to provide them with reliable information about the relative effectiveness of instructional strategies for promoting conceptual change, evidence must be provided to show that this instrument is both internally consistent and able to justify the inferences drawn from the test. These two properties are known as reliability and validity.

7.1 Reliability

Reliability is the extent to which a test is repeatable and yields consistent scores (Crocker & Algina 1986). There are several types of reliability, many of which involve multiple administrations of the same test over time. However, in some situations, such as with the LSCI, the test will be administered only once (only the posttest is used for estimating reliability). Testing instruments like the LSCI are not only concerned with how students score on the specific test items per se but are also intended to generalize to a larger content domain of possible items that might have been asked. One way to estimate how students' performance on the test can be generalized is to determine how consistently students performed across items or subsets of items on the given test (Crocker & Algina). In these situations, reliability is estimated using internal consistency methods.

A standard measurement of internal consistency for dichotomously scored tests is Cronbach's alpha. This reliability index corresponds to the lower bound of a test's internal consistency. The formula for Cronbach's alpha is shown below, where k = number of test items, sigma(i) squared = variance of item i, and sigma(x) squared = variance of entire test.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right)$$

The value of alpha represents the smallest fraction of total score variance that is due to true score variance (rather than errors in measurement). The minimum acceptable value for alpha is typically set at 0.7 (Litwin 1995), such that less than 30% of the score variance is due to measurement errors. The postinstruction value of Cronbach's alpha for the LSCI multi-institution field test was calculated to be a satisfactory 0.77. For comparison, the reliability indices for field tests of the Astronomy Diagnostic Test (ADT) and the Lunar Phases Concept Inventory (LPCI) were 0.77 and 0.68, respectively (Deming 2002; Lindell 2001).

When examinees perform consistently across items within a test, the test is said to have item homogeneity. Test items are considered homogeneous if they represent the same content domain such as the one defined in Section 2, Purpose and Specifications. If items on a test are representative of the topics covered by the course that a test aims to evaluate, students are more likely to perform consistently across all items than if the test is composed of items from a broad range of disciplines. Internal consistency, and therefore test reliability, is also lower if the items constituting the test are poorly written or are not of approximately equal relevance to the class (i.e., if some items are addressed as a main theme of a course, whereas others are taken from an obscure textbook reference). Responses to a reliable test differ because the examinees have conflicting opinions about the correct answer, not because the questions have multiple or ambiguous interpretations or are irrelevant to the material covered by the test.

7.2 Validity

Test validation is the procedure by which evidence is gathered to determine if test items satisfactorily represent a concept domain and whether the test measures the properties that it proposes to measure (Crocker & Algina 1986). Three types of validity were explored for the LSCI: face validity, content validity, and concurrent validity.

7.2.1 Face Validity

A test that appears to be a good measure of students' understanding of the concepts addressed by the test is said to have face validity (Trochim 2005). The question to be answered in determining the face validity of the LSCI was, therefore, "Do the questions composing the LSCI appear to represent the content domain most appropriate for measuring introductory college astronomy students' understanding of light and spectroscopy?" A thorough review of textbooks, surveys of introductory college astronomy faculty, and syllabi (Slater & Adams 2003; Slater et al. 2001; Zeilik & Morris-Dueer 2005) strongly suggests that the LSCI does indeed have face validity. Information gathered through student interviews and colleagues' reviews of the inventory supports this claim of face validity.

7.2.2 Content Validity

Content validation is an evaluation of the scientific accuracy of the items constituting a test (Trochim 2005). This characteristic was explored by having a team of experts review the LSCI for the relevance of the content with respect to the intended test population of introductory college astronomy students; for whether the items adequately sample the defined content domain; and for consensus on the correct answer to each item. Of the panel of eight experts consulted, which included a mixture of astronomy instructors and education researchers, concerns were expressed most frequently over two of the items also flagged for review based on questionable values of difficulty or discrimination (items 2 and 25). But overall, the experts exhibited an excellent level of agreement and therefore a satisfactory content validation result. The flagged items will be discussed further in an upcoming paper detailing the results of two semesters of field testing (Bardar 2007).

7.2.3 Concurrent Validity

Concurrent validity is a measure of whether a test is able to successfully distinguish between populations as it has proposed to do (Trochim 2005). For the LSCI, concurrent validation was performed through analysis of whether the LSCI was able to measure a statistically significant difference in postinstruction scores on the LSCI between courses employing active engagement strategies, and those using traditional lecture-dominated methods. Tests of statistical significance (t-tests) were able to differentiate courses employing learner-centered techniques from courses relying heavily on lecture as the primary mode of instruction (Bardar 2006). More details of this result will be presented in an upcoming paper (Bardar 2007).

8. THE LSCI

The final field-tested and validated 26-item LSCI can be viewed and downloaded in PDF by clicking on the link below:

http://aer.noao.edu/auth/LSCIspring2006.pdf

9. SUMMARY

This article has outlined the process by which the LSCI was constructed, guided by the principles of classical test theory. Item analysis following a multi-institution field test revealed several candidate items for revision or elimination from future versions of the inventory. Calculation of the inventory's reliability index using Cronbach's alpha strongly suggests that the LSCI has an acceptable level of internal consistency and can be considered to be reliable in this respect. Additional data will determine whether flagged items should be eliminated from future administrations of the test. A thorough review of textbooks and faculty surveys, combined with interview results and peer review, indicates that the LSCI exhibits an acceptable level of face validity. Agreement by experts regarding the correct responses to test items and the appropriateness of the content domain for the intended population of introductory college astronomy students suggests that the LSCI demonstrates sufficient content validity to be used in assessing introductory college astronomy students' conceptual understanding of topics related to light and spectroscopy. Validation of the LSCI as a tool for evaluating the relative effectiveness of instructional techniques is explored elsewhere (Bardar 2006, 2007) through comparative statistical analysis of LSCI

results for courses implementing active engagement and traditional lecture strategies.

The methodology of the approach taken to develop the LSCI and establish its reliability and validity can be applied to creating additional inventories for other popular astronomy topics such as gravity, cosmology, or size and scale of the universe. A complete suite of concept inventories would allow instructors to explore student understanding of the major astronomical concepts most directly related to their course and serve as a useful tool in assessing instructional techniques across many content areas.

References

Allen, M. J., & Yen, W. M. 1979, *Introduction to Measurement Theory*, Long Grove, IL: Waveland Press.

Ambrose, B. S., Heron, P. R. L., Vokos, S., & McDermott, L. C. 1999, "Student Understanding of Light as an Electromagnetic Wave: Relating the Formalism to Physical Phenomena," *American Journal of Physics*, 67(10), 891.

Ambrose, B. S., Shaffer, P. S., Steinberg, R. N., & McDermott, L. C. 1999, "An Investigation of Student Understanding of Single-Slit Diffraction and Double-Slit Interference," *American Journal of Physics*, 67(2), 146.

Bardar, E. M. 2006, "Development and Analysis of Spectroscopic Learning Tools and the Light and Spectroscopy Concept Inventory for Introductory College Astronomy," Unpublished Doctoral Dissertation, Boston University, Boston, MA.

Bardar, E. M. 2007, "First Results from the Light and Spectroscopy Concept Inventory", in preparation.

Bardar, E. M. (Weeks), Prather, E. E., Brecher, K., & Slater, T. F. 2005, "The Need for a Light and Spectroscopy Concept Inventory for Assessing Innovations in Introductory Astronomy Survey Courses," *Astronomy Education Review*, 4(2), 20. http://aer.noao.edu/cgi-bin/article.pl?id=160.

Brecher, K. 1991, "Do Atoms Really 'Emit' Absorption Lines?" *The Physics Teacher*, 29, 454.

Comins, N. F. 2001, *Heavenly Errors*, New York: Columbia University Press.

Crocker, L., & Algina, J. 1986, *Introduction to Classical and Modern Test Theory*, New York: Holt.

Deming, G. L. 2002, "Results from the Astronomy Diagnostic Test National Project," *Astronomy Education Review*, 1(1), 52. http://aer.noao.edu/cgi-bin/article.pl?id=5.

Ebel, R. L. 1965, *Measuring Educational Achievement*, Englewood Cliffs, NJ: Prentice Hall.

Hake, R. R. 1998, "Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics*, 66, 64.

Haladyna, T. M. 1994, *Developing and Validating Multiple-Choice Test Items*, Hillsdale, NJ: Erlbaum.

Hufnagel, B. 2002, "Development of the Astronomy Diagnostic Test," *Astronomy Education Review*, 1(1), 47. http://aer.noao.edu/cgi-bin/article.pl?id=4.

Langley, D., Ronen, M., & Eylon, B. 1997, "Light Propagation and Visual Patterns: Preinstruction Learners' Conceptions," *Journal of Research in Science Teaching*, 34, 399.

Lee, S. 2002, "Students' Understanding of Spectra," Unpublished Doctoral Dissertation, Kansas State University, Manhattan.

Lindell, R. S. 2001, "Enhancing College Students' Understanding of Lunar Phases," Unpublished Doctoral Dissertation, University of Nebraska, Lincoln.

Litwin, M. S. 1995, *How to Measure Survey Reliability and Validity*, Thousand Oaks, CA: Sage Publications.

Mazur, E. 1997, *Peer Instruction*, Upper Saddle River, NJ: Prentice Hall.

Nunnally, J. C. 1967, *Psychometric Theory*, New York: McGraw-Hill.

Prather, E. E., Slater, T. F., Adams, J. P., Bailey, J. M., Jones, L. V., & Dostal, J. A. 2004, "Research on a Lecture-Tutorial Approach to Teaching Introductory Astronomy for Non-Science Majors," *Astronomy Education Review*, 3(2), 122. http://aer.noao.edu/cgi-bin/article.pl?id=132.

Sadler, P. M. 1998, "Psychometric Models of Student Conceptions in Science: Reconciling Qualitative Studies and Distractor-Driven Assessment Instruments," *Journal of Research in Science Teaching*, 35, 265.

Schwab, J. J. 1963, *Biology Teachers' Handbook*, New York: Wiley.

Slater, T. F., & Adams, J. P. 2003, *Learner-Centered Astronomy: Strategies for Teaching ASTRO 101*, Upper Saddle River, NJ: Prentice Hall.

Slater, T. F., Adams, J. P., Brissenden, G., & Duncan, D. 2001, "What Topics Are Taught in Introductory Astronomy Courses?" *The Physics Teacher*, 39, 52.

Tamir, P. 1971, "An Alternative Approach to the Construction of Multiple Choice Test Items," *Journal of Biological Education*, 3, 305.

Trochim, W. M. 2005, *The Research Methods Knowledge Base* (2nd ed.), http://trochim.human.cornell.edu/kb/index.htm.

Wosilait, K., Heron, P. R. L., Shaffer, P. S., & McDermott, L. C. 1999, "Addressing Student Difficulties in Applying a Wave Model to the Interference and Diffraction of Light," *Physics Education Research*, *American Journal of Physics* supplement, 67(7), S5.

Zeilik, M., & Morris-Dueer, V. J. 2005, "What Are Essential Concepts in 'Astronomy 101'?: A New Approach to Find Consensus from Two Different Samples of Instructors," *Astronomy Education Review*, 2(3), 61. http://aer.noao.edu/cgi-bin/article.pl?id=121.

Zeilik, M., Schau, C., & Mattern, N. 1998, "Misconceptions and Their Change in University-Level Astronomy Courses," *The Physics Teacher*, 36, 12.